Modelli Lineari Generalizzati: un'applicazione

Fedi Flavia

6 luglio 2006

Capitolo 1

Introduzione

L'obbiettivo di questo studio è quello di utilizzare la classe dei modelli lineari generalizzati (GLM) per modellare un insieme di dati di natura biomedica. In particolare si vogliono osservare le relazioni che intercorrono tra le espressioni di una malattia mitocondriale nota come malattia di Leber (atrofia ottica di Leber) ed alcuni aspetti legati alla biogenesi mitocondriale[1].
Prima di introdurre la classe dei GLM verrà introdotta brevemente nel primo capitolo la classe dei modelli lineari di cui i GLM rappresentano un'estensione. Il secondo capitolo sarà dedicato ad alcuni aspetti peculiari dei GLM mentre nel terzo capitolo verrà effettuata l'analisi dei dati. Una trattazione completa riguardante la classe dei GLM si può vedere in Azzalini(2002) nel sesto capitolo.

1.1 Modelli Lineari

L'obbiettivo dei modelli lineari è quello di studiare la relazione che intercorre tra le variabili che caratterizzano un fenomeno.

In particolare si assume che una *variabile risposta*, chiamata Y, sia legata linearmente ad una o più *variabili esplicative* fissate, chiamate $X_1, \ldots, X_k$. La dipendenza lineare di Y rispetto alle variabili esplicative viene introdotta assumendo che la media della variabile di risposta sia una combinazione lineare delle variabili esplicative con $\beta_1, ..., \beta_k$ incogniti. Il valore osservato della variabile risposta è quindi composto da due termini, ovvero

$$Y = r\left(x_1, \ldots, x_n\right) + \epsilon = \sum_{j=1}^{k} \beta_j x_j + \epsilon \tag{1.1}$$

in cui il termine $r\left(x_1, \ldots, x_n\right)$, che rappresenta la combinazione lineare delle variabili esplicative, è detta componente *sistematica*, mentre ϵ è detta componente *accidentale* o di *errore*. Quest'ultima rappresenta lo scostamento di

[1] Insiemi di meccanismi che stanno alla base della proliferazione di mitocondri e della duplicazione del DNA mitocondriale

natura casuale tra Y e $r(x_1, \ldots, x_n)$.
Poichè $\mu = E(Y)$ è dato solo dalla componente sistematica dovrà essere $E(\epsilon) = 0$.
Supponiamo ora di essere in presenza di n osservazioni $Y_1, \ldots, Y_n$ e assumiamo che

$$Y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

con $i = 1, \ldots, n$ e dove Y_i è la $i-ma$ componente della variabile di risposta, x_{i1} rappresenta l'$i-mo$ valore della variabile esplicativa x_1, e così via. Si può riscrivere tutto in forma compatta come:

$$Y = X\beta + \epsilon$$

dove $X = (x_{ij})$ è detta **matrice di regressione** mentre i β sono chiamati **parametri di regressione**. In questa classe di modelli, detti modelli lineari, si assume che $\mathbb{E}[\epsilon] = 0$ e $Var[\epsilon]=\sigma^2\mathbb{I}$. Assumendo inoltre che $Y \sim N_n[X\beta, \sigma^2\mathbb{I}]$ il modello viene detto lineare normale.

1.2 Carenze dei Modelli Lineari

I modelli lineari presentano delle carenze, facciamo alcuni esempi

1. Può succedere che la relazione sia del tipo (1.1) ma con $r(\cdot)$ decisamente non lineare nei parametri.

2. Anche quando non si sa come è fatta $r(\cdot)$, dalla natura del fenomeno si è in grado di escludere a priori una relazione di tipo lineare. Ad esempio può succedere che il campo di variazione di $E(Y)$ non sia $(-\infty, \infty)$ come invece viene ipotizzato nel modello (1.1)

3. La varianza del termine di errore e quindi anche della variabile risposta è stata posta costante (ipotesi del secondo ordine) mentre spesso si riscontra empiricamente che ciò non è vero.

4. Nei modelli lineari si assume che la distribuzione della variabile risposta sia normale (ipotesi di normalità) ma spesso non si ha a che fare con variabili di questa natura. Si trasforma la variabile risposta in una nuova variabile con distribuzione normale. Questo metodo però non si può applicare sempre; in particolare quando Y è una variabile discreta può risultare problematico o anche impossibile trasfomare Y in modo da ottenere una variabile normale.

1.3 Un sottoinsieme della Famiglia Esponenziale

Introduciamo ora una classe di distribuzioni di probabilità che permetterà di generalizzare l'assunzione di normalità tipica dei modelli lineari.

Tale categoria di distribuzioni di probabilità è un sottoinsieme delle famiglie esponenziali che, ricordiamo, hanno densità

$$f(y) = q(y) \exp\left\{ \sum \mu_i(\theta)\, t_i(y) - \tau(\theta) \right\}. \tag{1.2}$$

Data l'importanza che tale classe di distribuzioni di probabilità avrà nel nostro studio, ne introduciamo una notazione specifica. Per una variabile continua reale Y diremo che

$$Y \sim EF\left(b(\theta), \frac{\psi}{\omega} \right)$$

se Y ha funzione di densità del tipo

$$f(y) = \exp\left\{ \left(\frac{\omega}{\psi} \left(y\theta - b(\theta) \right) \right) + c(y, \psi) \right\} \tag{1.3}$$

dove θ e ψ sono dei parametri scalari ignoti, ω è una costante nota , e $b(\cdot)$ e $c(\cdot)$ sono funzioni note la cui scelta individua una particolare distribuzione di probabilità. Per ogni particolare scelta di ψ, che è detto *parametro di dispersione*, la (1.3) costituisce una famiglia esponenziale di parametro θ, ma bisogna sottolineare che la (1.3) non sempre appartiene alla famiglia esponenziale nel caso in cui sia θ che ψ varino simultaneamente. Si consideri per ora ψ fissato. La (1.3) la si può scrivere come

$$\begin{aligned} f(y) &= \exp\left\{ \left(\frac{\theta}{\psi} y\omega - b(\theta) \frac{\omega}{\psi} + c(y; \psi) \right) \right\} && (1.4) \\ &= \exp\left\{ \left(\frac{\theta}{\psi} \omega y - b(\theta) \frac{\omega}{\psi} \right) \right\} \exp\left\{ c(y; \psi) \right\} && (1.5) \end{aligned}$$

Quindi facendo un parallelo con la (1.2) con i=1 otteniamo che

$$\begin{aligned} t(y) &= y && (1.6) \\ \mu(\theta) &= \frac{\theta}{\psi}\omega && (1.7) \\ q(y) &= \exp\left\{ c(y; \psi) \right\} && (1.8) \\ \tau(\theta) &= b(\theta)\frac{\omega}{\psi} && (1.9) \end{aligned}$$

Si vuole calcolare la media e la varianza di Y.
Per potelo fare dobbiamo per prima cosa determilarle nel caso di una famiglia esponeziale.
La $f(y)$ essendo una distribuzione di probabilità avrà:

$$\int f(y, \theta) dy = 1,$$

inoltre $f(\cdot)$ è sufficientemente regolare da poter giustificare la derivazione sotto segno di integrale e quindi che:

$$0 = \frac{d}{d\theta}1 = \frac{d}{d\theta}\int_y f(y;\theta)dy \quad = \quad \int_y \frac{d}{d\theta}f(y;\theta)dy.$$

Osserviamo allora che la derivata prima della mia $f(\cdot)$ è

$$\begin{aligned}\frac{d}{d\theta}q(y)exp\left(\mu(\theta)t(y)-\tau(\theta)\right) &= q(y)\left(\mu'(\theta)t(y)-\tau'(\theta)\right)exp\left(\mu(\theta)t(y)-\tau(\theta)\right)\\ &= f(y;\theta)\left(\mu'(\theta)t(y)-\tau'(\theta)\right)\end{aligned}$$

Quindi si ottiene che

$$\int_y f(y,\theta)\left((\mu'(\theta)t(y)-\tau'(\theta)\right)dy \quad = \quad 0$$

Essendo $\mu'(\theta)$ e $\tau'(\theta)$ costanti all'interno degli integrali si avrà

$$\mu'(\theta)\int_y f(y;\theta)t(y)dy = \tau'(\theta)\int_y f(y;\theta)dy$$

In definitiva :

$$\mathbb{E}\left\{t(y)\right\} = \frac{\tau'(\theta)}{\mu'(\theta)} \tag{1.10}$$

Per il calcolo della varianza osserviamo prima di tutto che

$$0 = \int_y \frac{d}{d\theta}\left[f(y,\theta)\left((\mu'(\theta)t(y)-\tau'(\theta)\right)\right]dy \tag{1.11}$$

dove

$$\begin{aligned}f'(y;\theta) &= q(y)\left(\mu'(\theta)t(y)-\tau'(\theta)\right)exp\left(\mu(\theta)-\tau(\theta)\right)\\ &= f(y;\theta)\left(\mu'(\theta)t(y)-\tau'(\theta)\right)\end{aligned}$$

ora si può calcolare esplicitamente la (1.11)

$$\begin{aligned}0 &= \int_y f'(y;\theta)\left(\mu'(\theta)t(y)-\tau'(\theta)\right)+f(y;\theta)\left(\mu''(\theta)t(y)-\tau''(\theta)\right)dy\\ &= \int_y f(y;\theta)\left(\mu'(\theta)t(y)-\tau'(\theta)\right)^2+f(y,\theta)\left(\mu''(\theta)t(y)-\tau''(\theta)\right)dy\end{aligned}$$

L'integrale del primo addendo è il momento secondo della variabile continua $\{\mu'(\theta)t(y)-\tau'(\theta)\}$; tale variabile ha valor medio nullo, in base alle

relazioni ottenute precedentemente e quindi il momento secondo coincide con la varianza si può quindi scrivere

$$\begin{aligned} Var\left\{\mu'(\theta)t(y)-\tau'(\theta)\right\} &= -\int_y f(y)\left(\mu''(\theta)t(y)-\tau''(\theta)\right)dy \quad (1.12)\\ &= -\mathbb{E}\left\{\mu''(\theta)t(y)-\tau''(\theta)\right\} \quad (1.13)\\ &= \frac{\mu'(\theta)\tau''(\theta)-\mu''(\theta)\tau'(\theta)}{\mu'(\theta)} \quad (1.14)\\ & \quad (1.15)\end{aligned}$$

e in definitiva si ottiene

$$Var\left\{t(y)\right\} = \frac{\mu'(\theta)\tau''(\theta)-\mu''(\theta)\tau'(\theta)}{\mu'(\theta)^3}. \quad (1.16)$$

A questo punto osservando che nella formula (1.3) la statistica t(y) assume in realtà proprio il valore y si può concludere che se $Y \sim EF(b(\theta), \frac{\psi}{\omega})$

$$\mathbb{E}\left(y\right) = \frac{\tau'(\theta)}{\mu'(\theta)} = b'(\theta)$$

$$Var\left(y\right) = \frac{\mu'(\theta)\tau''(\theta)-\mu''(\theta)\tau'(\theta)}{\mu'(\theta)^3} = b''(\theta)\frac{\psi}{\omega}$$

che è quello che si voleva determinare. Procedendo allo stesso modo con derivate di ordine superiore si possono ottenere espressione dei momenti superiori della v.c Y. Si vede peraltro come la (1.16), e ancor più le espressioni dei momenti superiori, si semplificano notevolmente se $\mu(\theta) = \theta$, si parla in tal caso di parametrizzazione naturale della famiglia esponenziale.
Se Y è una ossevazione da una v.c $Gamma(\nu, \frac{\nu}{\lambda})$, il parametro naturale, come vedremo nel capitolo terzo, è rappresentato dal $-\nu^{-1}$.
Concludiamo osservando che per le successive elaborazioni algebriche si porrà

$$\begin{aligned} \mu &= b'(\theta) \quad (1.17)\\ V(\mu) &= b''(\theta)|_{\theta=b'^{-1}(\mu)} \quad (1.18)\end{aligned}$$

Capitolo 2

Modelli Lineari Generalizzati

2.1 Dai modelli lineari ai modelli lineari generalizzati

I **Modelli Lineari Generalizzati** sono dei modelli che includono quelli lineari e che sono una naturale estensione di essi. Si considera il caso in cui la funzione $r(\cdot)$, introdotta nella formula (1) non sia lineare e le variabili non siano normali. Questa nuova classe di modelli non è molto ampia da un punto di vista strettamente matematico ma è sufficientemente flessibile da incorporare un gran numero di situazioni rilevanti per le applicazioni pratiche. Inoltre la classe dei MLG ha anche il pregio di permettere una trattazione unificata di una serie di modelli specifici, che prima dell'introduzione di questa classe erano trattati come casi singoli e non come sottocasi di un modello generale.

Per definire questa nuova classe di modelli si riconsiderano gli elementi caratteristici dei modelli lineari.
Per la generica unità i-esima poniamo $\eta_i = x_i^T \beta$ dove x_i^T è la i-esima riga della matrice X per $i = 1, \ldots, n$. Tale quantità incognita verrà anche chiamta *predittore lineare*.
Ricordando che nei modelli lineari normali veniva ipotizzato che $Y_i \sim N\left(\mu; \sigma^2\right)$ dove la relazione tra μ_i e il predittore lineare η_i era rappresentata dalla funzione identità, la classe dei MLG si ottiene estendendo la formulazione precedente in due direzioni:

1. si pone Y_i non strettemente Normale ma $Y_i \sim EF\left(b(\theta_i), \frac{\psi}{\omega_i}\right)$ tale che $b'(\theta_i) = \mu_i$;
2. si prendono in considerazione altre forme di legame tra il predittore lineare η_i e il valor medio μ_i, ovvero si ipotizza $g(\mu_i) = \eta_i$.

Sintetizzando il tutto si afferma che un MLG è caratterizzato dai seguenti

elementi:

$$
\begin{aligned}
Y_i &\sim EF\left(b(\theta_i), \frac{\psi}{\omega_i}\right) && (2.1)\\
g(\mu_i) &= \eta_i && (2.2)\\
\eta_i &= x_i^T \beta && (2.3)\\
& && (2.4)
\end{aligned}
$$

con $b'(\theta_i) = \mu_i$. Più analiticamente un MLG è caratterizzato dai seguenti elementi:

1. le osservazioni $y_1, \ldots, y_n$ sono tratte da variabili continue $Y_1, \ldots, Y_n$ tra loro indipendenti;
2. ciascuna Y_i ha distribuzione del tipo $EF\left(b(\theta_i), \frac{\psi}{\omega_i}\right)$ con $\mathbb{E}(Y_i) = \mu_i = b'(\theta_i)$ per $i = 1, \ldots, n$;
3. esiste una funzione $g(\cdot)$ tale per cui $g(\mu_i) = x_i^T \beta$,dove x_i è un vettore di costanti e β un vettore di parametri;
4. le funzioni $g(\mu), b(\theta)$ e $c(y; \psi)$ e il parametro di dispersione ψ sono comuni a tutte le Y_i, mentre il fattore peso ω può variare da individuo a individuo.

2.2 Verosimiglianza e Informazione di Fisher

Date le osservazioni campionarie $y_1, \ldots, y_n$ si vuole procedere a fare inferenza sui parametri β e ψ con particolare interesse per β poichè determina la relazione tra le variabili esplicative e la media μ.
Sia p la dimensione di β e $X = (x_{ij})$ la matrice $n \times p$ con i-esima riga ${x_i}^T$.
Essendo tutte le componenti indipendenti, si ha che la log-verosimiglianza è

$$
\begin{aligned}
l(\beta) &= \sum_{i=1}^{n} \left(\frac{\omega_i \left(y_i \theta_i - b(\theta_i)\right)}{\psi} + c_i(y_i, \psi) \right)\\
&= \sum_{i=1}^{n} l_i(\beta)
\end{aligned}
$$

Per ottenere le equazioni di verosiglianza si calcola

$$
\begin{aligned}
\frac{\partial l_i}{\partial \beta_j} &= \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} && (2.5)\\
\frac{\partial l_i}{\partial \theta_i} &= \frac{y_i - b'(\theta_i)}{\frac{\psi}{\omega_i}} && (2.6)\\
& && (2.7)
\end{aligned}
$$

ed essendo $\mu = b'(\theta)$ e $Var(Y_i) = b''(\theta)\frac{\psi}{\omega_i}$

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{\omega_i}{\psi} Var(Y_i)$$

e poichè $\eta = x_i^T \beta$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

In definitiva quindi le equazioni della verosimiglianza per β sono

$$\sum_{i=1}^{n} \frac{(y_i - \eta_i) x_{ij}}{Var\{Y_i\}} \frac{\partial \mu_i}{\partial \eta_i} = 0 \tag{2.8}$$

Per ottenere l'informazione de Fisher si considerano le derivate seconde di $l_i(\beta)$ ottenendo

$$\begin{aligned}
-\mathbb{E}\left\{\frac{\partial^2 l}{\partial \beta_i \partial \beta_k}\right\} &= \mathbb{E}\left\{\frac{\partial l_i}{\partial \beta_j}\frac{\partial l_i}{\partial \beta_k}\right\} \\
&= \mathbb{E}\left\{\left(\frac{(Y_i - \mu_i)\, x_{ij}}{var\{Y_i\}}\frac{\partial \mu_i}{\partial \eta_i}\right)\left(\frac{(Y_i - \mu_i)\, x_{ik}}{var\{Y_i\}}\frac{\partial \mu_i}{\partial \eta_i}\right)\right\} \\
&= \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{x_{ij}x_{ik}}{Var\{Y_i\}^2}\mathbb{E}\left\{(Y_i - \mu_i)^2\right\} \\
&= \frac{x_{ij}x_{ik}}{Var\{Y_i\}}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2
\end{aligned}$$

Quindi la matrice di informazione attesa ha elemento (j,k)-esimo

$$\sum_{i=1}^{n} \mathbb{E}\left\{\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k}\right\}$$

ovvero in notazione matriciale

$$I(\beta) = X^T \widetilde{\mathcal{W}} X$$

dove

$$\widetilde{\mathcal{W}} = \begin{pmatrix} \widetilde{w_1} & 0 & \cdots & 0 \\ 0 & \widetilde{w_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \widetilde{w_n} \end{pmatrix}$$

avendo posto

$$\widetilde{w_i} = \frac{1}{var\{Y_i\}}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

2.3 Legami Canonici e Statistiche Sufficienti

La $g(\mu)$ si può scegliere in modo arbitrario in quanto non è soggetta a particolari restrizioni.
Ad esempio si può scegliere in modo tale che θ_i coincida con η_i, parametro naturale della famiglia esponenziale ovvero:

$$g(\mu_i) = \theta_i \tag{2.9}$$

questo particolare legame è detto *legame canonico*. In questo caso si verifica che

$$\begin{aligned} l(\beta) &= \frac{1}{\psi}\left\{\sum_{i=1}^{n}\omega_i\left(y_i\theta_i - b\left(\theta_i\right)\right)\right\} + \sum_{i=1}^{n} c_i(y_i,\psi) && (2.10)\\ &= \frac{1}{\psi}\left\{\sum_{i=1}^{n}\omega_i\left(y_i {x_i}^T\beta - b({x_i}^T\beta)\right)\right\} + \sum_{i=1}^{n} c_i(y_i,\psi) && (2.11)\\ &= \frac{1}{\psi}\left\{\left(\sum_{i=1}^{n}\omega_i y_i x_i\right)^T\beta - \sum_{i=1}^{n}\omega_i b\left(x_i^T\beta\right)\right\} + \sum_{i=1}^{n} c_i\left(y_i,\psi\right) && (2.12)\\ & && (2.13) \end{aligned}$$

che mostra che $(\sum_i \omega_i y_i x_i)$ è una statistica sufficiente per β nel caso che il parametro ψ sia assente oppure noto.
Se ψ è ignoto, ma la verosimiglianza è ancora distribuita esponenzialmente, la $(\sum_i \omega_i y_i x_i)$ è comunque una componente della statistica sufficiante minimale. Avendo posto $g(\mu_i) = \theta_i$ dà luogo anche ad altri vantaggi .
Per quanto riguarda le derivate della log-verosimiglianza si ha che

$$\frac{d\mu_i}{d\eta_i} = \frac{d\mu_i}{d\theta_i} = \frac{db'(\theta_i)}{d\theta_i} = b''(\theta_i)$$

tenendo presente la (2.9), e quindi

$$\frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)x_{ij}}{var\{Y_i\}} b''(\theta_i) = \frac{\omega_i(y_i - \mu_i)x_{ij}}{\psi} \tag{2.14}$$

questo semplifica la (2.8) ed inoltre implica

$$\sum_i y_i x_{ij} \;=\; \sum_i x_{ij}\hat{\mu_i}$$

nel caso comune che $\omega_i = 1$
In notazione matriciale si scrive

$$X^T y = X^T \hat{\mu}$$

indicando con $\hat{\mu}_i$ i valori di μ_i corrispondenti alla SMV $\hat{\beta}$ di β Si supponga che una colonna di X sia I_n allora $X^T y = X^T \hat{\mu}$ ci dice che il totale dei valori osservati y è uguale al totale dei valori interpolati $\hat{\mu}$ e un'uguaglianza analoga vale per le altre colonne di X.
Per quando riguarda invece l'informazione di Fischer derivando la (2.14) si ottiene

$$\frac{\partial^2 l}{\partial \beta_i \beta_k} = \mathbb{E}\left\{\frac{\partial^2 l}{\partial \beta_i \beta_k}\right\}$$

E quindi l'informazione attesa e osservata coincidono.

Riportiamo una tabella in cui, tra le altre cose sono rappresentati i legami canonici relative ad alcune distribuzioni.

Distribuzione	Normale $N(\mu, \sigma^2)$	Binomiale/m $Bin(m, \mu)/m$	Gamma $G(\omega, \omega/\mu)$
Supporto	$(-\infty, \infty)$	$\{0, 1/m, ..., 1\}$	$(0, \infty)$
ψ	σ^2	1	ω^-1
ω	1	m	1
$b(\theta)$	$\theta^2/2$	$\log\left(1 + e^\theta\right)$	$-\log(-\theta)$
$c(y;\psi)$	$-\frac{1}{2}\left(\frac{y^2}{\psi} + \log(2\pi\psi)\right)$	$\log\begin{pmatrix} m \\ my \end{pmatrix}$	$\frac{\log(\omega y)^\omega}{\log y \log \Gamma(\omega)}$
$\mu(\theta)$	θ	$e^\theta/1 + e^\theta$	$-\frac{1}{\theta}$
Legame canonico	Identità	logit	reciproco
$V(\mu)$	1	$\mu(1-\mu)$	μ^2

Tabella 2.1: `Elementi caratteristici di alcune distribuzioni`

2.4 Stima del parametro di dispersione

Ricordando che $\eta_i = x_i^T \beta_i$ e quindi $\hat{\eta}_i = x_i^T \hat{\beta}$ con $\mu_i = \eta_i$ in definitiva si ottiene che $g(\hat{\mu}_i) = \eta_i$ e una volta calcolato $\hat{\beta}$, si ha a disposizione il vettore dei valori medi stimati $(\hat{\mu_1}, ..., \hat{\mu_n})$ (medie stimate) e quindi ponendo la funzione g l'identità si ha che

$$\tilde{\psi} = \frac{1}{n-p} \sum_i \frac{\omega_i (y_i - \hat{\mu_i})^2}{V(\hat{\mu_i})}$$

che è basato sulla relazione $VarY_i = V(\mu_i)\frac{\psi}{\omega_i}$.
Nel caso dei modelli lineari con legame identità $\tilde{\psi} = s^2 = \hat{\sigma}^2 \frac{n}{n-p}$ ecco perchè mettiamo n-p e non n.

2.5 Adeguatezza dei Modelli

2.5.1 Devianza

Si considera il problema di confrontare due MLG. Siano M_1 e M_2 due modelli distinti con la condizione che $M_2 \subset M_1$, si tratta quindi di *modelli annidati*. In particolare si prendono due modelli lineari generalizzati, e su M_2 si impongono dei vincoli supplementari sui parametri del predittore lineare. Ovvero se M_1 è un modello contenente p_1 parametri e M_2 un modello con p_2 parametri, si impongono dei vincoli del tipo $g_i(\beta) = 0$ per $i = 1, ..., p_2 - p_1$
Come metodo per confrontare i modelli annidati si utilizza quello del rapporto di massima verosimiglianza. Nei modelli lineari la seguente quantità

$$Q\left(\hat{\beta}\right) = ||y - \hat{\mu}||^2$$

viene chiamata *Devianza* e si può dimostrare che il il test del rapporto di verosimiglianza è funzione della devianza associata a ciascuno dei modelli.
Infatti la verosimiglianza dipende dai dati solo attraverso D, come si vede scrivendo

$$l\left(\hat{\beta}\right) = c - \frac{n}{2}\log\left(\sigma^2\right) - \frac{D}{2\sigma^2}.$$

Si chiama *modello saturo* quello in cui le stime di μ_i cincidono con y_i, cosa che è possibile con un modello contenente n parametri.
Tale modello non è di utilità pratica, ma serve come termine di confronto per il modello effettivamente in esame. Tecnicamente esso serve a liberare la log-verosimiglianza dalle costanti arbitratrie.
Se si confrontasse la log-verosimiglianza del modello in questione con quello saturo, con $\tilde{\mu_i} = y_i$, si ottiene che il rapporto di verosimiglianza tra i due modelli, saturo e annidato, vale

$$-2\left\{l\left(\hat{\beta}\right) - l\left(\tilde{\beta}\right)\right\} = \frac{D}{\sigma^2}$$

ed è pari alla devianza stessa, a meno di una costante. Inoltre per confrontare due modelli annidati, il rapporto di verosimiglianza diventa

$$W = -2\left\{l\left(\hat{\beta}_2\right) - l\left(\hat{\beta}_1\right)\right\} = \frac{D_2 - D_1}{\sigma^2}$$

e cio è sostanzialmente la differenza delle devianze a meno di una costante σ^2 che abbiamo supposto nota.
Si può ora estendere il concetto di devianza nell'ambito del MLG. Il modello saturo rimane invariato e si indica con $\tilde{\theta_i}$ il corrispondente valore di θ_i.
Risulta allora che

$$W(y) \quad = \quad -2\left[l\left(\hat{\beta}\right) - l\left(\tilde{\beta}\right)\right] \tag{2.15}$$

$$= -2\sum_i \frac{\omega_i}{\psi}\left[\left(y_i\hat{\theta}_i - b\left(\hat{\theta}_i\right)\right) - \left(y_i\tilde{\theta}_i - b\left(\tilde{\theta}_i\right)\right)\right] \quad (2.16)$$

$$= \frac{\sum_i d_i}{\psi} \quad (2.17)$$

Sia $D\left(y;\hat{\mu}\right) := \sum_i d_i$ la *devianza di un* MLG dove il generico termine d_i è il contributo della i-esima osservazione alla devianza; la quantità $D(y;\hat{\mu})/\psi$ è detta *devianza normalizzata*. Nel caso dei modelli annidati

$$\frac{D\left(y;\hat{\mu_2}\right) - D\left(y;\hat{\mu_1}\right)}{\psi} \overset{d}{\longmapsto} \chi^2_{p_1-p_2}$$

per $n \to \infty$ e se p_1 e p_2 sono costanti rispetto a n

2.5.2 Residui

I residui costituiscono uno strumento grazie al quale si può valutare informalmente l'adeguatezza di un modello lineare.
Si estende il concetto di residuo agli MLG nel modo che segue. Si prendono in esame

$$e_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V\left(\mu_i\right)/\omega_i}}$$

che è detto *residuo di Pearson*

$$e_i^D = sign(y_i - \hat{\mu})\sqrt{d_i}$$

che prende il nome di *residuo di devianza*

$$e_i^A = \frac{A(y_i) - A(\mu_i)}{A'(\mu_i)\sqrt{V\left(\mu_i\right)}}$$

in cui

$$A(x) = \int \frac{1}{V(x)^{1/3}}dx$$

è scelta in modo tale da rendere la loro distribuzione normale.

Capitolo 3

Un'applicazione in campo biomedico

3.1 Descrizione dell'insieme di dati

In questo capitolo verrà presentata una applicazione reale in cui la classe dei modelli lineari generalizzati viene utilizzata per modellare un insieme di dati di natura biomedica.
Abbiamo infatti a disposizione un campione di 134 persone, tutte appartenenti ad una famiglia residente in Brasile in cui una malattia genetica nota come malattia di Leber si trasmette di generazione in generazione. Su ogni unità sono state rilevate le seguenti variabili

1. Y = quantità media di DNA mitocondriale rilevate in tre misurazioni differenti
2. X_1 = età
3. X_2 = sesso
4. X_3 = espressione della malattia

La nostra popolazione è quindi composta da individui di sesso maschile o femminile, che codificheremo con sesso1 e sesso0, ognuno dei quali può essere sano, malato o portatore della malattia. Tali stati verranno indicati con stato2, stato0, stato1.

Sesso e Stato sono delle variabili qualitative mentre l'età di ogni individuo e la media del numero di copie di DNAM sono quantitative. Scopo dell'analisi è studiare la relazione che intercorre tra la quantità di DNA mitocondriale e le altre variabili. In particolare dato che il supporto della variabile risposta Y è il semi-asse positivo, assumiamo che tale variabile Y

abbia distribuzione di tipo gamma che ha densità

$$\begin{aligned} f(y;\nu,\lambda) &= \frac{\lambda^{\nu} y^{\nu-1} e^{-\lambda y}}{\Gamma(\nu)} \\ &= \exp\{\nu log(\lambda) + (\nu-1) log(y) - \lambda y - log(\Gamma(\nu))\} \\ &= \exp\left\{\nu\left(log(\lambda) + \left(\frac{\nu-1}{\nu}\right) log(y) - \frac{\lambda}{\nu} y - \frac{log(\Gamma(\nu))}{\nu}\right)\right\} \end{aligned}$$

Ponendo $\theta = -\frac{\lambda}{\nu}$ abbiamo che la densità della v.a Gamma può essere riscritta nel modo seguente

$$\begin{aligned} f(y;\nu,\lambda) &= \exp\left\{\nu\left(log(-\theta\nu) + \left(\frac{\nu-1}{\nu}\right) log(y) + \theta y - \frac{log(\Gamma(\nu))}{\nu}\right)\right\} \\ &= \exp\{\nu(\theta y + log(-\theta\nu)) + (\nu-1) log(y) - log(\Gamma(\nu))\} \\ &= \exp\{\nu(\theta y + log(-\theta) + log(\nu)) + (\nu-1) log(y) - log(\Gamma(\nu))\} \\ &= \exp\{\nu(\theta y + log(-\theta)) + \nu log(\nu) + \nu log(y) - log(y) - log(\Gamma(\nu))\} \\ &= \exp\{\nu(\theta y + log(-\theta)) + \nu log(\nu y) - log(y) - log(\Gamma(\nu))\} \end{aligned}$$

per cui ponendo

$$\begin{aligned} \psi &= \nu^{-1} \\ \omega &= 1 \\ b(\theta) &= -log(-\theta) \\ c(y;\psi) &= \nu log(\nu y) - log(y) - log(\Gamma(\nu)) \end{aligned}$$

risulta che il nostro modello può essere scritto nel modo seguente

$$Y_i \sim EF(-log(-\theta_i), \psi)$$

.

Per la nostra applicazione abbiamo scelto un link di tipo logaritmico[1] e la variabile stato è stata opportunamente dicotomizzata. La relazione che in base al nostro modello lega il valore atteso della quantità mitocondriale dell'i-esimo soggetto rispetto alle covariate è dunque la seguente:

$$\log(E[Y_i]) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}$$

dove β_1 rappresenta l'intercetta della della relazione tra il logaritmo del valore atteso di Y e l'età nella popolazione dei malati di sesso femminile, x_{i2} rappresenta l'indicatore dello stato portatore della malattia, x_{i3} rappresenta

[1] per link si intende $g(\mu_i) = x_i^T \beta$ ovvero quella funzione che esplicita la relazione tra predittore lineare e il valore atteso della distribuzione

l'indicatore dello stato assenza della malattia, x_{i4} indica l'età e x_{i5} l'indicatore dello stato uomo.
Come primo obbiettivo stimiamo i coefficienti β. Tali parametri, infatti, definiscono la relazione che intercorre tra il numero di DNAM e le covariate. Tra i vari pacchetti statistici che permettono di stimare un modello lineare generalizzato abbiamo scelto di utilizzare (`www.R-project.org`). In particolare avendo chiamato (`dati`) la matrice le cui colonne rappresentano le variabili descritte in precedenza, possiamo stimare il modello richiesto attraverso la sintassi

```
glm(dna.medie ~ stato+età+sesso,family=Gamma,link=log,data=tele)
```

la formula

`dna.medie`$\sim$`stato+età+sesso`

indica che si vuole ottenere delle informazioni sul numero di DNAM in relazione allo stato all'età e al sesso del paziente.
L'output di R è il seguente

	estimate	std.error	t value	Pr($>\|t\|$)
(intercept)	6.388547	0.145934	43.777	>2e-16
stato1	-0.473373	0.114015	-4.152	<0.0001
stato2	-1.077437	0.083164	-12.956	0
età	-0.001839	0.002684	-0.685	0.495
sesso1	0.047045	0.079504	0.592	0.555

Sulla colonna Estimate sono riportati i valori di $\hat{\beta}$ ovvero le stime del coefficiente di regressione.

I dati relativi all'intercetta sono riferiti alla popolazione di riferimento che è costituita dalle ragazze malate. Il valore 6.38 indica quindi che il logaritmo del numero medio di copie di DNAM in questa popolazione quando l'età assume il valore zero è appunto 6.38. Il valore riferito allo Stato1 e $\hat{\beta}$=-0.473373 indica che se un paziente assume stato 1, ovvero è portatore della malattia, ha un valore medio di DNAM inferiore di 0.4 rispetto alla mia popolazione di riferimento.
Il valore riferito allo stato2 $\hat{\beta}$=-1.077437 indica che se un paziente assume stato 2, ovvero è sano ha un valore medio di DNAM inferiore di 1.4 rispetto alla mia popolazione di riferimento.
Il valore riferito al sesso1 $\hat{\beta}$=0,047045 indica che un ragazzo ha un valore medio di DNAM superiore di 0.04 rispetto alla mia popolazione di riferimento.

L'output di R riporta anche i risultati del test di ipotesi $H_0 : \beta_i = 0$ contro $H_1 : \beta_i \neq 0$ sui coefficienti di regressione. Tali risultati sono riportati sulla colonna chiamata `t value` e nell'ultima colonna vengono riportati i valori p. Ricordiamo che se il valore p è piccolo tenderemo ad accettare l'ipotesi alternativa rifiutando quindi l'ipotesi nulla $H_0 : \beta_i = 0$. Analizzando l'ultima colonna possiamo quindi concludere che i coefficienti di regressione relativi alle variabili età e sesso possono essere ipotizzati uguali a zero. Ricordiamo infine che il test di Wald misura $|\hat{\beta}_i - \beta_i|$ opportunamente standardizzato.
Un altro modo per valutare l'evidenza a favore di una data ipotesi riguardante un coefficiente di regressione è valutare contemporaneamente la stima $\hat{\beta}_i$ e il corrispettivo standard error. Prendiamo in considerazione tali valori relativi alla variabile sesso. In questo caso abbiamo ottenuto $\hat{\beta}_i = 0.04$ e il corrispettivo standard error è 0.07. Attraverso questi valori possiamo costruire un intervallo di confidenza allo 0.5% dato approsimativamente da $\hat{\beta}_i \pm 2$ standard error ovvero, nel nostro caso (-0.03,0.12). Poichè tale intervallo comprende lo zero possiamo accettare l'ipotesi che il coefficiente di regressione relativo al sesso sia nullo. Lo stesso risultato lo abbiamo ottenuto analizzando semplicemente il valore dell'ultima colonna ovvero 0,555 che, essendo alto, ci porta ad accettare $\beta = 0$. Quindi possiamo affermare che gli uomini rispetto alle donne malate hanno un numero di DNAM maggiore in media e su scala logaritmica di 0.047045 . Tutto quello che abbiamo visto fino ad ora poteva essere dedotto analizzando il grafico riportato nella figura 3.1. Tale grafico può essere interpretato come una matrice dove al posto ij viene rappresentato la variabile i-esima in relazione alla variabile j-esima. In particolare a noi interessa la prima riga in cui il numero di dnam è rappresentato graficamente in relazione alla variabile sesso, età e stato.
Notiamo ad esempio che il grafico di posto 1,2 ci porta ad affermare il numero di dnam non dipende in modo influente dal sesso mentre, se analizziamo l'elemento 1,4, vediamo che la variabile stato influisce in modo significativo. Ciò è conforme con quanto detto sopra infatti il numero dnam in relazione allo stato0 è superiore rispetto agli altri stati.

Analizziamo ora la devianza del modello che se piccola è indice di un buon adattamento ai dati. Ricordiamo che la devianza normalizzata è data da

$$D_N = 2\sum_{j=1}^{n} \left\{ logf\left(y_j; \tilde{\eta}_j, \hat{\psi}\right) - logf\left(y_j; \hat{\eta}_j, \hat{\psi}\right) \right\}$$

mentre la devianza è data da

$$D = \hat{\psi} \left\{ 2\sum_{j=1}^{n} \left[logf\left(y_j; \tilde{\eta}_j, \hat{\psi}\right) - logf\left(y_j; \hat{\eta}_j, \hat{\psi}\right) \right] \right\}$$

Nel nostro caso abbiamo che la devianza è 38.382.
Dall'output di R ricaviamo anche che la devianza normalizzata risulta 12.758.

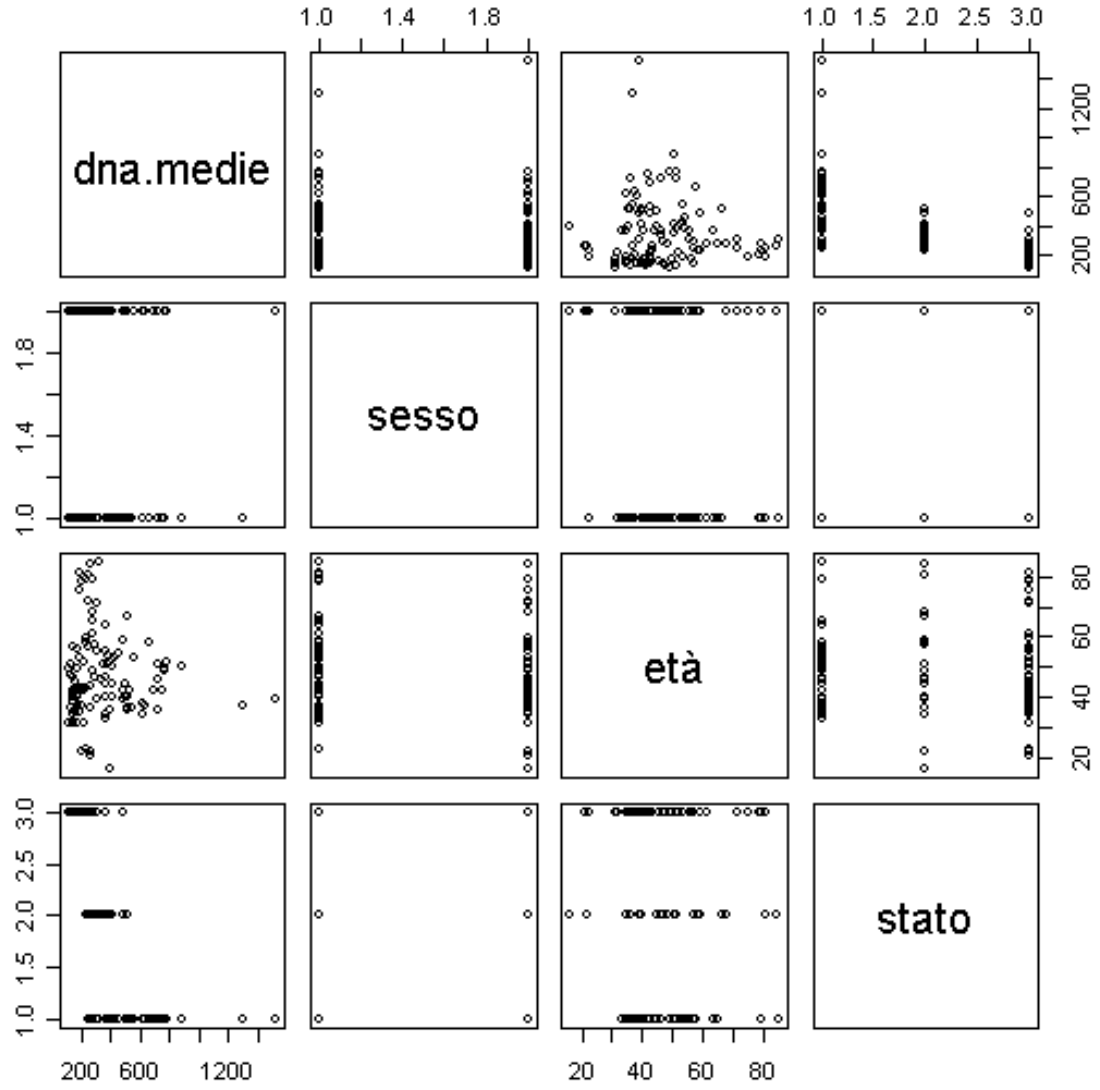

Figura 3.1: Rappresentazione grafica dei dati relativi alla malattia di Leber

Osserviamo inoltre che, prendendo un modello in cui la variabile dipendente `dnam` viene supposta dipendere solo dalla variabile `stato`, la devianza normalizzata risulta 12.892. Essendo la diminuzione nella devianza normalizzata dovuto all'introduzione delle variabil `sesso` e `età` pari a 12.892-12.758=0.134 (ovvero di scarsa entità) possiamo concludere che queste ultime due variabili potrebbero essere non inserite nella costruzione del modello. Ovvero il modello ottimale per l'analisi di questi dati è non è quello da noi studiato ma un modello in cui l'unica variabile esplicativa è la variabile espressione della malattia.

Bibliografia

[1] A.Azzalini. *Inferenza Statistica: una presentazione basata sul concetto di verosimiglianza*

Indice